Touching BASE: Connecting ESGF to HPSS

Sasha Ames[1], **Sam Fries**[1], Alex Sim[2], Dean Williams[1]

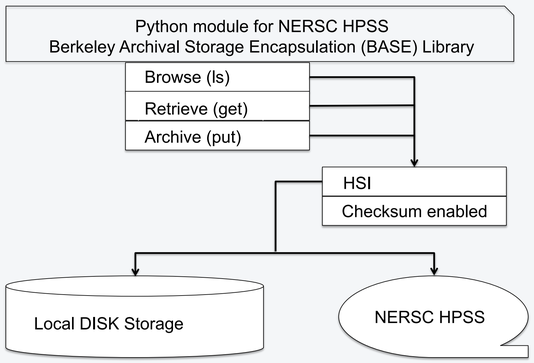
[1]AIMS Group, Lawrence Livermore National Laboratory
[2]SDM Group, Lawrence Berkeley National Laboratory

Abstract

Accessing data stored on tape archives is difficult, time consuming, and prone to error. The ACME project plans to create hundreds of terabytes to petabytes of data, which exceeds the capacity of disk-based archives. To address this, we are bridging HPSS and ESGF, allowing data sets stored on tapes to be accessed through the same methods that climate scientists are already familiar with. LBNL's Berkeley Archival Storage Encapsulation (BASE) library provides a simple Python API for retrieving metadata as well as actual data from HPSS. We are creating a Python web application that uses BASE to access and retrieve data, and allow that data to be published to ESGF. Our initial platform will test HPSS at NERSC with ESGF nodes at LLNL, with plans to deploy at other ACME sites.

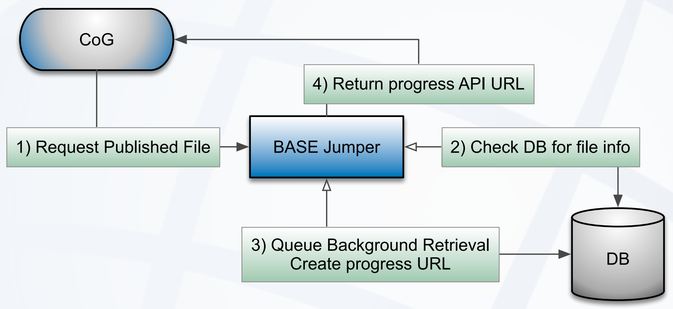
BASE

The Berkeley Archival Storage Encapsulation Library comes from the experience of building, maintaining, and operating the BeStMan and HRM systems from 1998 to 2015. It is a small Python module that provides three main functions (Browsing, retrieving, and archiving). Under the covers, it leverages HSI for performing these tasks, creating a greatly simplified design.



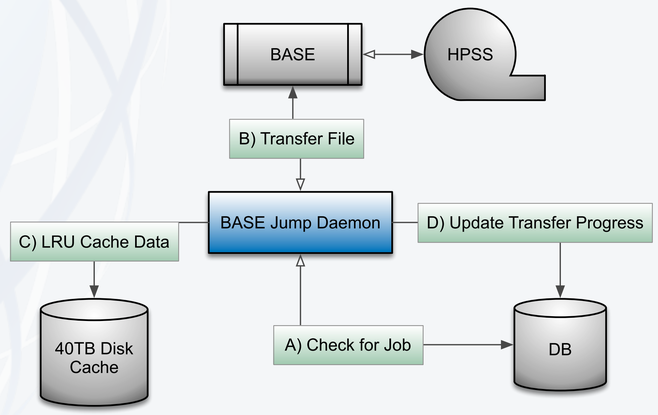
BASE Jumper

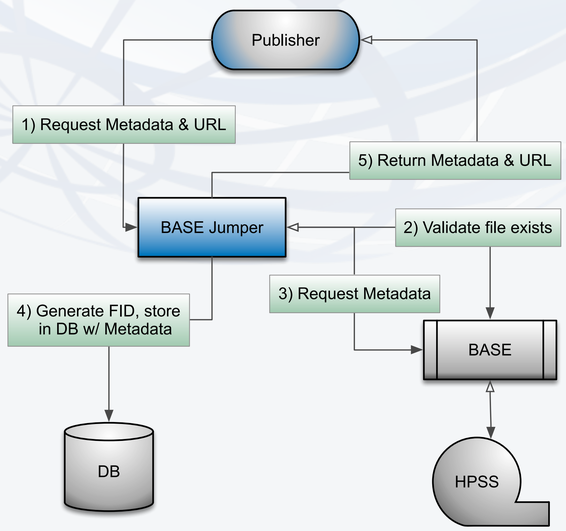
BASE Jumper will be the application that connects ESGF and HPSS. It will consist of two components; a web application that manages authentication and interaction with consumers of the service, and a daemon that manages retrieval and storage of actual data.



The web application will be integrated with CoG to provide a user-friendly view into the operations of the daemon. An indicator for data published via BASE Jumper will be added to the views in CoG, with a note indicating that access will be asynchronous. When a user requests a file from CoG that has been published from HPSS, BASE Jumper will check the user's access to the file and queue the job to processed by the daemon. A shared database will connect the web application and the daemon, allowing progress updates to be shared back to CoG and the user via an API call to the web application.

The daemon will manage data retrieval from HPSS using BASE. It will manage a local cache of data transferred off HPSS, and evict files as space is needed for new transfers. As it receives data from HPSS, it will update the progress of the transfer in the database. Upon completion, a notification to the user will be triggered, with a guaranteed-availability download window and a URL for downloading the file.





The publisher will use API calls available through the BASE Jumper web application to extract metadata from data on HPSS. From the metadata, the publisher will produce URLs that reference the BASE jumper service, stored in the Postgres DB, TDS catalog and SOLR. These URLs will appear in the search results shown in CoG, from which users will gain access to each file or data set stored in HPSS.

This work is supported by the ESGF consortium: DOE, NASA, NOAA, IS-ENES, NCI. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DEAC52-07NA27344.